Introduction

This Annex lists the data-fields available from the Airwave Study Tissue Bank treatments and medications export.

Document Configuration

Title	Data Dictionary, Annex E
Subject	Treatments & Medications Export
Version	1.0
Status	Draft
Author	Heard, Andy H, Database Manager at Imperial College London.
Published	26th January 2023
Filename	annex-e.docx

Medications Overview

The treatments and medications dataset is a single file, each record of which is a medication (or occasionally a treatment) that participants reported that they were taking at the time of their clinic visit. Although we asked participants and nurses to limit themselves to prescription medicines prescribed by a doctor, the dataset also contains non-prescription drugs available over the counter at pharmacies, and all kinds of self-help potions and supplements. We sought only the drug name, not dosage or other prescribing details. When participants forgot to bring this information with them, we recorded whatever they reported to us.

In this documentation and elsewhere, “treatments” and “medications” are synonymous.

Description of Data Cleaning

The names of treatments and medications were entered by the nurse either by clicking on one of a pre-defined pick-list of treatments, or by entering the medication as free-format text. For free-format medications, the nurse could also record additional notes, which sometimes explains the reason for it being taken, or the dosage, etc.

Nurses were instructed to choose from the pick-list in preference to entering free-format text, but this was not always followed. Participants occasionally entered a medication from the pick-list and then re-entered it into a free-format field, usually so that they can provide a descriptive note (pick-list items did not provide space for notes).

Reference Dataset

Items found in the picklist self-identify, of course, and these constitute c. 70% of the dataset. The remaining 30% of free-format prescriptions cannot always be simply matched to an item from a standard resource such as the British National Formulary (BNF). For these, we built an algorithm to “fuzzy-match” entered data against a set of treatments obtained from public-domain resources. In practice, this works well, with only 2.8% of reports being unmatched to a standard medication. There were also ambiguities, 1.5% of reports matching more than one possibility. For these cases, manual inspection often shows that the algorithm has matched a single substance spelled slightly differently for generic and branded versions. There is still work to do to clean up the remaining uncertainties.

We have not provided a description for the matching algorithm as it remains a work in progress; but feel free to ask the Study team if you are curious. Moreover, the reference set is neither complete nor up-to-date, and researchers should bear this in mind when using the data.

Caveat Emptor

We make no claims that our interpretation of the medications data is accurate or complete, and researchers should make their own decision before using this dataset. Conversely, researchers who are able to add to the quality and completeness of the treatments dataset are strongly encouraged to provide us with their additions and corrections for the wider benefit of the community.

Data Labels

The output file described below is tab-separated and includes a header record of column names.

Label	Data Type	Description
barcode	NUMBER (5)	Health-screening identifier.
med#	NUMBER(2)	A sequence number within barcode for each medication.
observed_value	VARCHAR2 (400)	This is either the label chosen from a fixed-list of predefined medications, or the narrative description entered by the nurse.
when_observed	DATE	When the medication was reported by the participant. This is *not* the prescription date.
field_type	VARCHAR2 (11)	One of two values according to the method by which the treatment was entered: **PICK LIST**: chosen from a pick-list of treatments. **FREE FORMAT**: entered as free format text.
matched_to	VARCHAR2 (500)	When field_type = **FREE FORMAT**, this is the medication within the reference set that we matched observed_value to. See also match#. When we cannot match observed_value, matched_to is empty.
match#	NUMBER(1)	A sequence number for each match matched_to. For example: "Aspirin or paracetamol" will flag as two medications. Researchers should decide for themselves how to interpret these results. Each match was scored, and a lower match# indicates a better score according to our algorithm.
notes	VARCHAR2 (500)	Any free-format text that was entered when field_type = **FREE FORMAT**.

Version History

VERSION	Filename	Date Exported	Total Rows	CRC-64
2	treatments-v2.tsv	26th January 2023	51,620	09D2CD8B5508EF49
1	med_exp_001.csv	1st November 2019	46,910	B99A59D284F637D7